Discovery of Classification Rules in Prediction of Applications Usage in Social Network Data (Facebook Application Data) Using Data Mining Algorithms

Dr. R. Geetha Ramani, P. Nancy

Associate Professor, Dept of IST
College of Engg, Anna University, Guindy Campus, Chennai, India.

Research Scholar, Dept of IST,
College of Engg, Anna University, Guindy Campus, Chennai, India.

Abstract— Data Mining is a process, which involves automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns and gathers knowledge which was hidden earlier. It involves various processes of which classification, Association rule mining and clustering gain major attention. One of the emerging application areas of Data Mining is Social Networks. The focus of the research is towards framing classification rules to predict the patterns in installation/ usage of Facebook applications towards the top most popularly installed/used application. The Dataset used in this research is Facebook Application installation / usage Dataset which contains details of installation of nearly 16,800 applications among 3 lakhs users. The work begins with Data Preprocessing where installation/usage of top 10 applications (selected based on the count of installations made by users) were used for Process. Various Data Mining Classification Algorithms such as RndTree, ID3, C-RT, CS-CRT, C4.5 and CS-MC4, Decision List, Naives Bayes are applied to preprocessed data individually and analyzed and the Classification rules for predicting the installation / usage of particular application are identified. The training Phase is processed with Training data and the testing phase is tested with test data.

Keywords – Data Mining; Algorithms; Applications; Social Network; Prediction; Facebook; error rates; Classification Rules.

I. Introduction

Data Mining makes use of data analysis tools to identify patterns and relationships in voluminous datasets. Data Mining Applications use classification, clustering, prediction, Association rule mining, pattern Recognition and Pattern Analysis. Data Mining has found its application in a variety of areas where Social Networks play a major role. Social network has become omnipresent in today's world. It paves way to share information among any number of people all over the world. Many Online Social networks exist, some of which include Orkut, Face book, FriENSTER, Myspace etc., Face book gains prominence over these by providing a record of maximum usage among users with almost 845 million active users as of February 2012. With more than 845 million active users around the world, Face book is today's most prominent social utility to connect with diverse audiences, including friends, family, co-workers, constituents, and consumers. These connections occur not just through Face book features but through applications ("apps") developed by third parties over Face book Platform.

A. Background of Facebook Applications

Facebook alone has over 81,000 third-party applications [5]. The Face book users install many applications through developer platforms. The Face book Developer Platform was launched in May 2007 [14] with little elaboration and only about eight applications in schedule. As months passed the Platform showed rapid growth with more than 35,000 applications by July 2008 [14]. The first step to create a Facebook application requires the developer to register the application with Facebook. Each application is assigned an application-id and a private application key. All communication between the application and Facebook's servers has to be signed with this key. A user can install an application by visiting the application's landing page, and accepting the dialog specifying the access rights of the application. However, the user can only accept or cancel the dialog. It is not possible to selectively grant or deny access to individual profile information.

This paper presents our research work on analysis of installation / usage of various Face book applications by the active users and reveals the best classification algorithm in classifying the usage of top ten applications. The classification rules obtained can be used in predicting the usage/installation of a particular application in future. This paper uses the Face book Application Dataset produced by Minas Gjoka and his team.

The original Dataset consists of two subdivisions. The first subdivision includes a data set that consists of data obtained from Adonomics [4], a service based on statistics reported by FB [15], for a period of 6 months from Sept. 2007 until Feb. 2008. It gives a detailed description of nearly 16,800 applications, the number of installations of each application and the number of users who use the application at least once during a day, called Daily Active Users (DAU). The second data set gives details of the Face book user profile with the various applications used by each user. Our work focuses on the second dataset for the purpose of finding classification towards in the installation/ usage of top ten applications.

B. Organization of the Paper

The rest of the paper is organized as follows. Section 2 reviews the related work in this area. Section 3 describes the data mining framework and the details of the dataset used in this research. It also briefs about the various classification algorithms that are applied to this dataset. Experimental results are discussed in Section 4 while Section 5 concludes the paper

II. Related Work

The work carried out so far by other researchers that are related to Facebook data is concisely presented here. However, we wish to state that no previous research has targeted the Facebook application dataset that we have used in our research.

Three Facebook applications were developed and launched which have achieved a combined subscription base of over 8 million users. Exploration of existence of 'communities', with high degree of interaction within a community and limited interaction outside the community within the context of Face book applications [12],[13].

Wei Panv and team proposed computational model to predict mobile application (known as "apps") installation using social networks and explained the challenges involved in their work. They show the importance of considering many factors in predicting app installations, and observed the surprising result that app installation was indeed predictable [18].

Online context allows studying social influence processes by tracking the popularity of a complete set of applications installed by the user population of a social networking site. This captures the behavior of all individuals who can influence each other in this context. By extending standard fluctuation scaling methods, the collective behavior induced were analyzed by 100 million application installations, and have revealed that two distinct regimes of behavior emerge in the system [16].

A Proxy on the Client Side system that provides a Facebook user with fine-grained access control capabilities over which parts of his / her private profile information can be accessed by third-party applications. [17].

D. E. Brown, V. Corruble, and C. L. Pittard [6] compared decision tree classifiers with back propagation neural networks for multimodal classification problems. J. Catlett [7] has explained how knowledge patterns can be generated from large databases. M. James [8] in his book describes the various classification algorithms. T. Cover and P. Hart [9] performed classification using K-NN and proved its accuracy.

III. Data Mining Framework

This section gives a brief description of the overall system design and Dataset used in this research. The overall design of the proposed system is given in Figure 1 and each of the components is addressed in further sections briefly. The design framework for the classification of Facebook Application usage/installation comprises of the training phase which incorporates the process of training data selection, data pre-processing and generation of classification rules through classification algorithms. This is followed by an Evaluation phase wherein the classifiers are evaluated based on their error rates. The Test phase verifies the chosen classifier's accuracy on classifying an unseen Application data

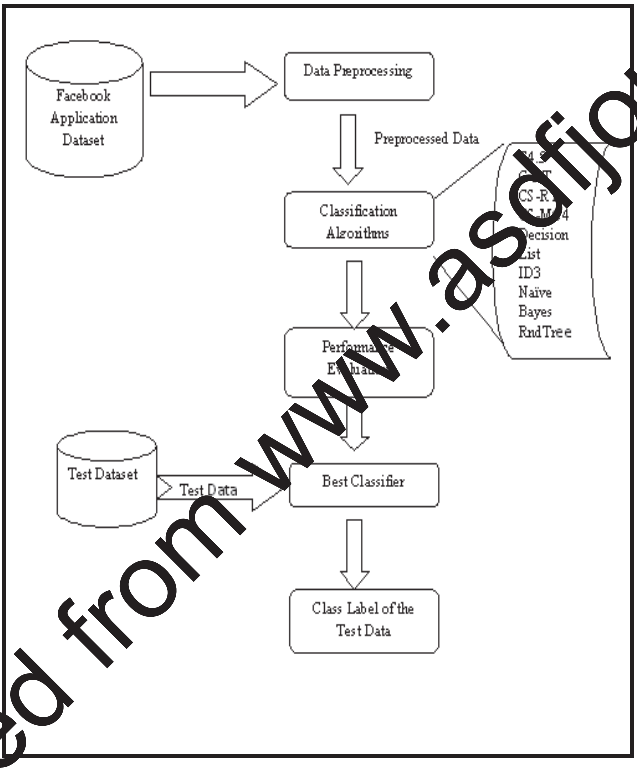


Figure 1. Overall System Design

A. Dataset Description

The Dataset utilized for this research is Face book Application dataset. The original Dataset consists of two subdivisions. The first data set consists of detailed description of nearly 16,800 applications, the number of installations/usage of each application and the number of users who use the application at least once during a day, called Daily Active Users (DAU). The second data set gives details of the Face book user profile with the various applications installed/used by each user. This work uses only the second dataset which contains a list of installed/used applications for 297K Face book users. UserIDs are anonymised. The Dataset is of the form

Table I Dataset Under Study

Data Source	Period	Data Element
FB User Profile	20/02/08 to 27/02/08	Users list, application installed

A sample dataset is shown in the Table 2.

Table II A Sample Dataset

User id	App1	App2	....	App773
1	2339854854	22800106120		5954997258
2	5902932866	8123226859		3361908998
3	6280837251	5737540558		5437153164
4	2363570816	2424357634		7029683973
5	17501549056	5902932866		6280837251

B. Data Preprocessing

The original Dataset shows the installation/usage of various applications by the users. The number of applications installed/used by a user ranges from 3 to 773. The data are preprocessed by identifying the top ten applications (based on the number of installations) among the users. This research work focus on exploiting information about the classification rules in predicting the usage of top ten Facebook applications

C. Classification Algorithms

The goal of Classification is to build a set of models that can correctly foresee the class of the different objects [3]. Classification Algorithms like RndTree, ID3, C-RT, CS-CRT, C4.5 and CS-MC4, Decision List, Naives Bayes were applied. The following are brief outline of some Classification Algorithms.

Rnd Tree Algorithm

The classification works as follows[2]: the Random Trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". The pseudo code of the Rnd Tree algorithm for this domain is given in Figure 2.

```
Begin {FT = {collection of all predictor features -forest}
IP = {input data – feature vector}
Repeat {
Compare the Attribute Values (av) of IP with FT.
If (IP.av == FT.av) then take the positive branch
Else take the negative branch        }
for all IP until leaf node is reached.
End
```

Figure 2. Rnd Tree Algorithm Pseudocode

ID3 (Iterative Dichotomiser) Algorithm

It is an Algorithm used to generate a decision tree invented by Ross Quinlan. ID3 is precursor to the C4.5 Algorithm. The work flow of the Algorithm is shown in Figure 3.

```
ID3 (Examples, Target_Attribute, Attributes)
Create a root node for the tree
If all examples are positive, Return the
single-node tree Root, with label = +.
If all examples are negative, Return the
 single-node tree Root, with label = -.
If number of predicting attributes is empty,
then Return the  single node tree Root,
with label = most common value of
the target attribute in the examples.
Otherwise Begin
A = The Attribute that best
Classifies examples.
Decision Tree attribute for Root = A. For each
possible value, $v_i$, of A,
Add a new tree branch below Root,
Corresponding to the test A = $v_i$.
Let Examples($v_i$) be the subset of examples
 that have the value $v_i$ for A If Examples($v_i$) is
empty
Then below this new branch add a
leaf node with label = most common target value in
the examples
Else below this new branch add the
sub tree ID3 (Examples($v_i$), Target Attribute,
Attributes – {A})
End  Return Root
```

Figure 3. ID3 Algorithm

C4.5 Algorithm

It is also called as statiscal classifier [2]. The pseudo code of the general Algorithm is as follows:

Check for base cases. For each attribute a, Find the normalized information gain from splitting on a. Let a_best be the attribute with the highest normalized information gain .Create a decision node that splits on a_best. Recurse on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

C-RT & CS-CRT

The CART method [2] under Tanagra is a very popular Classification tree learning algorithm. CART builds a decision tree by splitting the records at each node; according to the function of a single attribute it uses the gini index for determining the best split. The CS-CRT is similar to CART but with cost sensitive classification.

CS-MC4A

Cost sensitive decision tree Algorithm [2]. This version uses m-estimate smoothed probability estimation (a generalization of Laplace estimate). It minimizes the expected loss using misclassification cost matrix for the detection of the best prediction within leaves. The precondition required for this Algorithm is that at least one discrete attribute (target) and one or more discrete / continuous attribute (input) must be available

IV. Experimental Results

This section shows the analysis and results after executing various Classification Algorithms and explores the results of the same. The whole experiment is carried out with the Data Mining tool TANAGRA. The Applications are ranked based on the count of installations and top ten are ranked based on the count of installations and top ten applications are identified. Classification Algorithms like C4.5, C-RT, CS-RT, CS-MC4, Decision List, ID3, Naïve Bayes and RndTree were applied to the pre-processed Data. The Performance of these Algorithms is evaluated based on the error rates. The installation / usage of Top Ten Applications considered for the work is shown in Table 3. The error rates of various Classification Algorithms are shown in Table 4.

Table III List of Top Ten Applications Installed

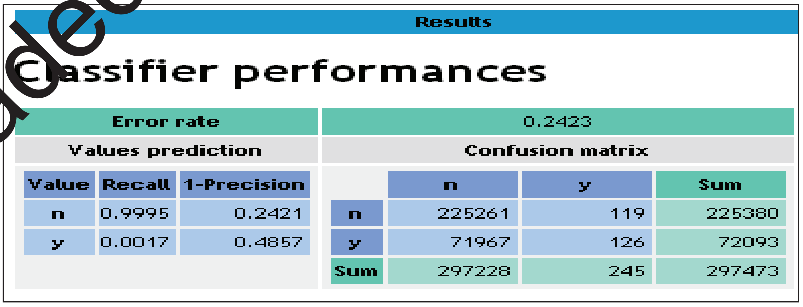
Application Number	Application Name	Number of Installations	Rank
2361831622	groups	253963	1
2305272732	Photos	190183	2
2601240224	Super Wall	108196	3
2425101550	Top Friends	101705	4
2386512837	Gifts	101053	5
2378983609	FunWall	99097	6
2357179312	rate amate	88826	7
2558160538	Movies	87609	8
2309869772	Poste items	72093	9
2345673396	hugme	69711	10

The performances of the Classification Algorithms were evaluated based on the error rates obtained.

Table IV Error Rates of Classification Algorithm for 10 Subsets

Top 10 Applications	Error rates of Various Classification Algorithm							
	C4.5	C-RT	CS-RT	CS_MC4	Decision List	ID3	Naïve Bayes	**Rnd Tree**
Groups	0.1463	0.1463	0.1463	0.1463	0.1463	0.1463	0.1463	**0.1463**
Photos	0.3369	0.3386	0.3386	0.3368	0.3518	0.3607	0.3485	**0.3368**
Super Wall	0.2245	0.2260	0.2260	0.2244	0.2350	0.2404	0.2303	**0.2244**
Top Friends	0.2753	0.2782	0.2782	0.2753	0.2828	0.2896	0.2836	**0.2750**
Gifts	0.3371	0.3382	0.3382	0.3370	0.3397	0.3397	0.3425	**0.3369**
FunWall	0.2270	0.2285	0.2285	0.2269	0.2404	0.2404	0.2363	**0.2269**
Rate amate	0.2260	0.2620	0.2620	0.2599	0.2660	0.2818	0.2792	**0.2599**
Movies	0.2611	0.2635	0.2635	0.2611	0.2710	0.2945	0.2734	**0.2610**
Posted items	0.2423	0.2424	0.2424	0.2423	0.2424	0.2424	0.2451	**0.2423**
Hug me	0.1960	0.1974	0.1974	0.1960	0.2117	0.2013	0.2175	**0.1967**

The error rates of various classification Algorithms are found using a confusion matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the System is confusing two classes (i.e. commonly mislabeling one as another). A sample Confusion Matrix for RndTree Algorithm is shown in Figure 4. Of all the Algorithms, RndTree Algorithm gave less error rates.



Results

Classifier performances

Error rate	0.2423

Values prediction

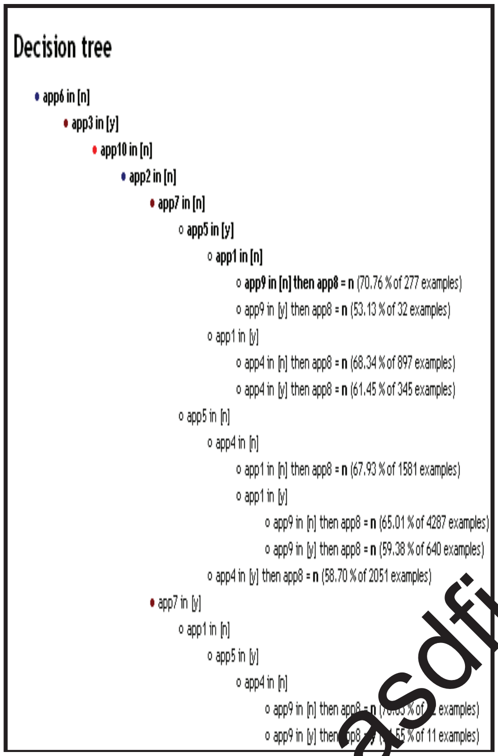
Value	Recall	1-Precision
n	0.9995	0.2421
y	0.0017	0.4857

Confusion matrix

	n	y	Sum
n	225261	119	225380
y	71967	126	72093
Sum	297228	245	297473

Figure 4. Confusion Matrix of RndTree Algorithm for Posted items Application

The rule generated by RndTree towards classification of installing/using application that is ranked 9 and ranked 1 is shown in Figure 5 and Figure 6.

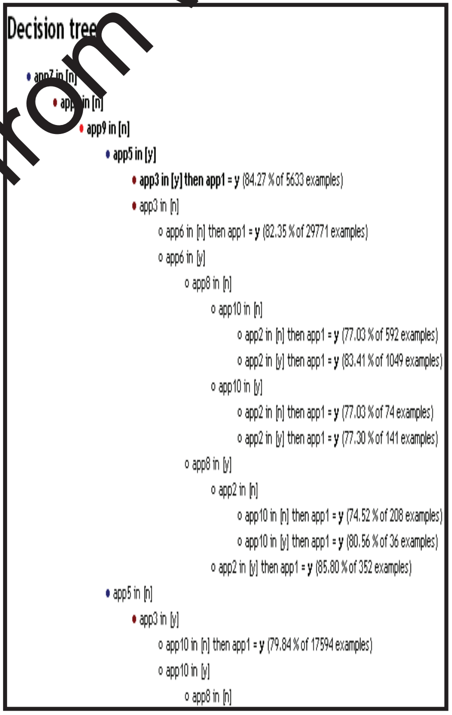


```
Decision tree

• app6 in [n]
    • app3 in [y]
        • app10 in [n]
            • app2 in [n]
                • app7 in [n]
                    ○ app5 in [y]
                        ○ app1 in [n]
                            ○ app9 in [n] then app8 = n (70.76 % of 277 examples)
                            ○ app9 in [y] then app8 = n (53.13 % of 32 examples)
                        ○ app1 in [y]
                            ○ app4 in [n] then app8 = n (68.34 % of 897 examples)
                            ○ app4 in [y] then app8 = n (61.45 % of 345 examples)
                    ○ app5 in [n]
                        ○ app4 in [n]
                            ○ app1 in [n] then app8 = n (67.93 % of 1581 examples)
                            ○ app1 in [y]
                                ○ app9 in [n] then app8 = n (65.01 % of 4287 examples)
                                ○ app9 in [y] then app8 = n (59.38 % of 640 examples)
                        ○ app4 in [y] then app8 = n (58.70 % of 2051 examples)
                • app7 in [y]
                    ○ app1 in [n]
                        ○ app5 in [y]
                            ○ app4 in [n]
                                ○ app9 in [n] then app8 = n ([illegible] % of [illegible] examples)
                                ○ app9 in [y] then [illegible] = [illegible] ([illegible].55 % of 11 examples)
```

Figure 5. A Snapshot of rule Generated by RndTree Algorithm For Posted items Application

The generated rules were also used to predict the installation/usage of intended applications and tested with test data and found to be correct.



```
Decision tree

• app7 in [n]
    • [illegible] in [n]
        • app9 in [n]
            • app5 in [y]
                • app3 in [y] then app1 = y (84.27 % of 5633 examples)
                • app3 in [n]
                    ○ app6 in [n] then app1 = y (82.35 % of 29771 examples)
                    ○ app6 in [y]
                        ○ app8 in [n]
                            ○ app10 in [n]
                                ○ app2 in [n] then app1 = y (77.03 % of 592 examples)
                                ○ app2 in [y] then app1 = y (83.41 % of 1049 examples)
                            ○ app10 in [y]
                                ○ app2 in [n] then app1 = y (77.03 % of 74 examples)
                                ○ app2 in [y] then app1 = y (77.30 % of 141 examples)
                        ○ app8 in [y]
                            ○ app2 in [n]
                                ○ app10 in [n] then app1 = y (74.52 % of 208 examples)
                                ○ app10 in [y] then app1 = y (80.56 % of 36 examples)
                            ○ app2 in [y] then app1 = y (85.80 % of 352 examples)
            • app5 in [n]
                • app3 in [y]
                    ○ app10 in [n] then app1 = y (79.84 % of 17594 examples)
                    ○ app10 in [y]
                        ○ app8 in [n]
```

Figure 6. A Snapshot of rule Generated by RndTree Algorithm For Group Application

V. Conclusion

Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. Social Network Data is vast and used in many researches. One such data “Facebook Application Dataset” is used in this research. There have been a large number of data mining Algorithms rooted in these fields to perform different data analysis tasks. In this paper, the comparisons on the performance of various Data Mining Classification Algorithms in effective prediction towards installation of top ten Facebook Applications were analysed. The classification rules produced by various Data Mining Classification Algorithms are evaluated based on the error rates. From the results it is clear that in all the top ten applications considered for the research RndTree Algorithm produced less error rates when compared to all other Algorithms and the rules generated by RndTree Algorithm predicted the installation/usage of intended application among users correctly. The accuracy is tested with a sample test data.

References

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, 1-36, Cambridge, 1996

[2] Tanagra Data Mining tutorials, http://data-mining- tutorials.blogspot.com/ this website provides detailed information on the basics of Data Mining Algorithms

[3] Dr. Varun Kumar, Luxmi Verma,” Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Algorithms in the Diagnosis of Breast Cancer” in IJCST Vol. 1, Issue 2, December 2010

[4] Adonomics. http://www.adonomics.com, May 2008

[5] Developer analytics. http://www.developeranalytics.com/,Apr 2008

[6] D. E. Brown, V. Corruble, and C. L. Pittard. A comparison of decision tree classifiers with back propagation neural networks for multimodal classification problems. Pattern Recognition, 26:953-961, 1993.

[7] J. Catlett. Mega induction: Machine Learning on Very large Databases. PHD Thesis, University of Sydney, 1991.

[8] M. James. Classification Algorithms. John Wiley, 1985.

[9] T. Cover and P. Hart. Nearest neighbor pattern classification. IEEE Trans. Information Theory, 13:21-27, 1967.

[10] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008-12-17.

[11] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996. Exclusive Ore Inc. The Exclusive Ore Internet Site, http://www.xore.com, 1999.

[12] Atif Nazir, Saqib Raza, Chen-Nee Chuah, 2008, Unveiling Facebook: A Measurement Study of Social Network Based Applications in IMC’08

[13] Atif Nazir, Saqib Raza, Chen-Nee Chuah, 2009, Network Level Footprints of Facebook Applications in IMC'09, November 4.–6, 2009, Chicago, Illinois, USA

[14] Facebook developer platform. http://developer.facebook.com/, Apr 2008.

[15] Facebook statistics, available at: http://www.Facebook.com/press/info.php?statistics

[16] Jukka-Pekka Onnela, Felix Reed-Tsochas,2010,Spontaneous emergence of social influence in online systems

[17] Manuel Egele_y, Andreas Moser_y, Christopher Kruegely, and Engin Kirda, 2011. PoX: Protecting Users from Malicious Facebook Applications in PERCOM Workshop

[18] Wei Panv, 2011 in Composite Social Network for predicting Mobile app Installation, AAAI.

[19] Minas Gjoka, Michael Sirivianos, Athina Markopoulou, Xiaowei Yang ,Poking Facebook: Characterization of OSN Applications in WOSN 2008.